Exploring human activity annotation using a privacy preserving 3D model

Mathias Ciliberto
Wearable Technologies,
Sensor Technology Research Centre,
University of Sussex
m.ciliberto@sussex.ac.uk

Francisco Javier Ordóñez
Wearable Technologies,
Sensor Technology Research Centre,
University of Sussex
f.ordonez-morales@sussex.ac.uk

Daniel Roggen
Wearable Technologies,
Sensor Technology Research Centre,
University of Sussex
daniel.roggen@ieee.org

Abstract
Annotating activity recognition datasets is a very time consuming process. Using lay annotators (e.g. using crowdsourcing) has been suggested to speed this up. However, this requires to preserve privacy of users and may preclude relying on video for annotation. We investigate to which extent using a 3D human model animated from the data of inertial sensors placed on the limbs allows for annotation of human activities. The animated model is shown to 6 people in a suite of tests in order to understand the accuracy of the labelling. We present the model and the dataset, then we present the experiments including the number of activities. We present 3 experiments where we investigate the use of a 3D model for i) activity segmentation, ii) for "open-ended" annotation where users freely describe the activity they see on screen, and iii) traditional annotation, where users pick one activity among a pre-defined list of activities. In the latter case, results show that users recognise with 56% accuracy when picking from 11 possible activities.

Author Keywords
Activity recognition; annotation; wearable technologies; 3D human model.

ACM Classification Keywords
H.5.m [Information interfaces and presentation]: Miscellaneous

Introduction

Activity recognition is fundamental for context-aware computing [3] and can be used to understand the habits of a user, to help patiences during their rehabilitation and more in general in the healtcare [2].

In order to achieve a reliable recognition of the activity of the user, a large annotated dataset is needed to train the machine learning classifier. Annotation requires that someone manually specify the actions carried out during the data recording. To do this, usually several videos are recorded jointly with the recording of inertial data. After the recording, the video from the cameras are synchronized with the data logged by the sensors, in order to annotate precisely each segment of data. This is a very time-consuming task, as address by authors in [10]. For this reason, it is usually done using cheap labour and/or crowdsourcing. In such a case, it is yet important to preserve the privacy of the user who recorded the data.

In this work we investigate to which extent a 3D human model animated directly from inertial sensors placed on user's limbs can be used to label the activities of that user while preserving his/her privacy.

The contribution of this paper are:

- a 3D human model created to reproduce the user movements. The model is developed in Java and it can be exported and deployed on many different platform, allowing a wide application in a crowdsourcing scenario.
- an investigation about the segmentation of the activities of the 3D model before annotating the data. We compare the results of segmentation done by the testers with the actual segmentation of activities in the dataset.
- an open-ended annotation test: during the experiment we let the user free to assign a custom label to each activity. The results of these test are presented using a tag cloud of the words used by testers.
- a comparison, using a confusion matrix, between the annotations chosen by the users using a set of possibilities and the true labels of each activity.

State of the art

The approaches to the annotation can be several and very diffent, as address in [4]. Usually the annotation of inertial data is done in post hoc using a video recorded together with them as help. It has been done by researchers in [10], which used a set of cameras to record the scene from three different perspectives. Then, using a custom developed software, which synchronized the videos with the inertial data. The activity of labelling is a tedious and very time-consuming task: as reported in that paper, for 30 minutes of recording, the annotation took 7-10 hours of analysis. In fact, during the labelling phase, the video synchronized with the inertial data requires that each sequence of data must be accurately analyzed to segment and recognize each activity correctly. For this reason, usually the annotation phase is done using cheap labour.

Recently, crowd-sourcing has been suggested to help reduce the cost or time for annotating datasets. is a process where a task can be completed by soliciting contributions from a large group of people. Thanks to this technique, the researchers can obtain a large amount of labelled data quickly. They can split the dataset in shorter segments and ask to other people to evaluate and annotate them. This can be obtained, for example, using tools like Amazon Mechanical Turk (MTurk) [1]: this is a webservice where users can ask for workforce. The workers can pick up a task and complete it earning a money reward. Crowdsourc-

ing has been used to tag human activity using the video, [8]. Crowdsourcing has also been used to label natural language [13], for speech recognition [9] and for multimedia tagging [12].

When the annotation task is done using the videos and this tools, one of main issue is to preserve the privacy of the subject in the dataset/video. To protect the anonymity of the subject in the dataset, the videos should be preprocessed. This preprocessing consist of apply a mask to the elements that in the video can be considered sizable from a privacy point of view. In [5] the authors use low resolution camera to preserve the privacy of the subject recorded. This step brings however additional time and work to the annotation task. Moreover, the preprocessing cannot be easily applied to all the element in order to do not alter too much the video source and to do not compromise the recognition of the activity.

A different approach can be the real-time annotation of the data: as shown in [6] and [11] it can be done using audio tag recorded by the subject of the dataset together with the inertial data. This method preserves the privacy and can be very accurate. It can be also used in a "open-ended" context thanks to the absence of a predefined set of labels. Real-time labelling requires the direct interaction of the user and in the everyday life this could be annoying.

Experimental setup

The model

The human model is developed using the open-source 3D engine called jMonkeyEngine [7]. This multiplatform engine written in Java allows to develop a human model from the ground up. It would be also possible to load an external model. We chose however to build our custom model to keep easier the handling of the animation and to guarantee more flexibility during the application of inertial data to each body part. As you can see in Figure 1, the model is a dummy build using some basic solids.

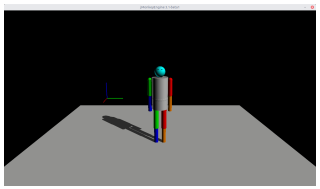


Figure 1: The human model

To animate the model, we used the data provided by 5 Inertial Measurement Units (IMUs) placed on the upper limbs and on the torso. The inertial data are espressed in quaternions. After a precomputation step, we applied the respective quaternion to each body part thanks to the engine, which handle directly this formalism. The multiplatform nature of the engine allows also to deploy this model and its animation in crowdsourcing scenario, because it can be easily integrated and used remotely, eg. in a webpage.

The engine allows to the users to rotate the camera around the model and to zoom in and out to better observe and evaluate each action.

The dataset

In this paper we use the the Opportunity Dataset [10] because it comprises a rich set of naturalistic activities. This dataset consist of inertial data about the absolute orientation of each limb during the session, which is recorded in a kitchen environment. These data have been recorded using a set of XSens MTx inertial sensors [14]. For our tests we used the "Drill" run subset. This set include data about 17 actions repeated consecutively for 20 minutes. Due to the absence of the environment in the 3D engine, we decide to join some of the similar labels (e.g. interacting with drawers at different heights is combined). "Open" and "Close" are considered diffent actions. From the initial 17 types of activities in the Opportunity dataset we obtain the following 11 types of activities which we aim to annotate in this paper:

- Open and close two different doors;
- Open and close three drawers at three different heights;
- Open and close a dishwasher;

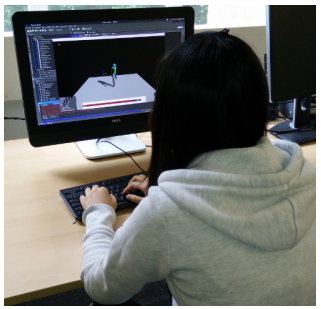


Figure 2: Setup of the experiment.

- Open and close a fridge;
- Clean a table;
- Drink from a cup;
- Toggle a switch.

Annotation experiments
We perfomed three experiments. The participants to the experiments were told that they would see a 3D model of a person performing typical activities in a kitchen. The participants were not given the list of activities at first. Essentially they have to "guess" from the animation of the model which activity may be undertaken. In the first, a 15 minutes animation is played by the model. During this animation the partipant must press the space-bar every time he/she notices something that they consider interesting and/or easy recognizable in the model movements. It is up to the tester to decide what is "interesting". This experiment is used to evaluate the capability of the users to segment the activities using the model. The interface used during this test is the same shown in Figure 1.

In the second and in the third experiment, a set of short animations of the body model are shown to the partecipants where the model performs exactly one of the 11 possible activities. For each of the 11 activities we showed the animations of 4 activity instances picked randomly from the dataset. In this ways, we show to the user a random but balanced set of activities. The set of 44 short animations has been showed to the tester in a random order. This set was different for each partecipats.

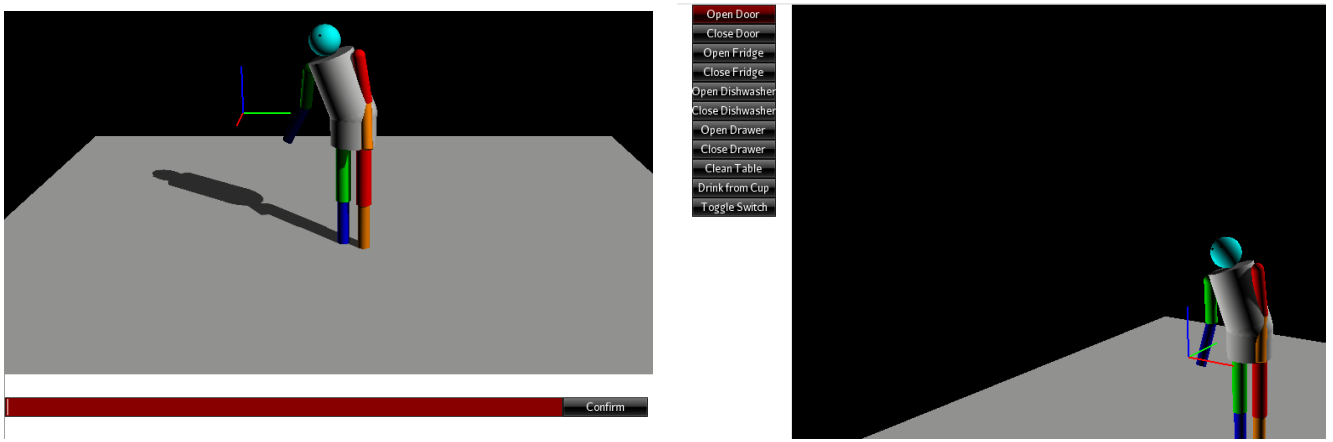
The second test is studied to investigate to which extent the application of our model in a "open ended" scenario. In this experiment, the task of the users is to insert a short label for each animation. We develop the interface displayed in Figure 3 in order to allow the user to enter the label. This interface showen up at the end of each animation of the set, and the user had not a time limit to enter the label. After he/she confirmed the inserted label, the next animation in the set is played.

The last experiment is useful to test the ability to annotate using a 3d model in a more common scenario. The system shows a push button for each of the 11 predefined activities. The participants must select which activity they think it was by pressing the corresponding push button with the mouse. The buttons are shown at the end of each short animation without any limit of time for the users. In Figure 3 are shown the buttons. After he/she selected a label, the next animation in the set is played. This corresponds to the common annotation approach where a pre-defined list of activities are annotated.

The experiment is made with 6 people, that are unaware about the dataset and the set of labels until the last test. The setup of the experiment is shown in Figure 2. All the participants deal with the tests in the same order and individually. They instructed before each test as to what they have to do next, in order to not influence the each phase of the experiment.

Results

Every experiment is needed to test a specific step or a different scenario of the annotation. With the first, we aim to evaluate the capability of the users to segment the activities. Figure 4 shows the results of the segmentation tests. The first row of the figure represents the distribution of the activities throughout the 15 minutes of the test. The next 6 rows shows the events pointed out by each partipant. Every vertical line is an event. It allows to compare the distribution of the event recorded by each tester and the actual distribution of the activities.



(a) **(b)**

Figure 3: Interfaces developed for second and third test.

As the experiment left the partipants free to decide what consider "interesting"; for this reason we observe a large variations in the frequency of the events recorded for each tester. This may be explained because some users tent to point out longer actions while other recorded more shorter task. An example of the first type of people is the User 3, who recorded less events than the User 5. However, it is possible to notice some similarities in the pattern of the events for some users and the actual segmentation pattern: the User 4 represents an example of this.

The second experiment aims to evaluate the application of our model in the "open-ended" scenario. In Figure 5 we show the tag clouds of the words entered to describe the activities by all users for each label. We noticed that most of the partecipants mistake the dishwasher with the oven: this is quite normal because both the appliances can have the same kind of door. Moreover, it is missing any rendering of the kitchen environment in the scene and no information is given to the user about the appliances and about the fornitures at this stage. However the testers correctly identify the difference between open and close the dishwasher.

In the last test, we investigate the common scenario where the user should annotate a dataset already segmented, choosing the correct label in a predefined "closed set". The results are presented in the Figure 6 using a confusion matrix between the choices of the users and the actual labels of the data.

It is observable as some activities are easy recognized by the user: "Drink from Cup" reaches an accuracy close to 100%. On the other hand, there are actions that are almost never identified: "Open the Fridge" is the task with the low-

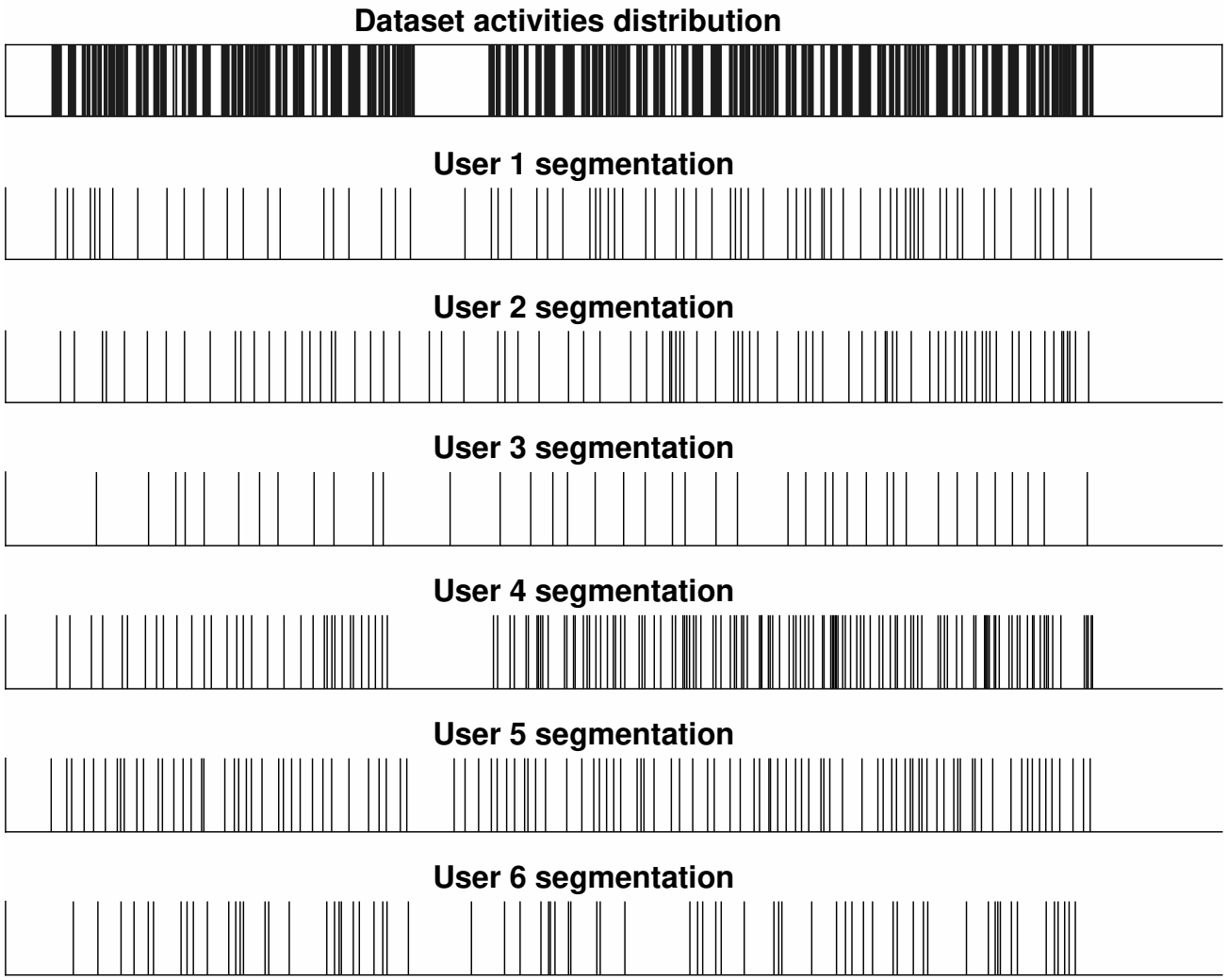
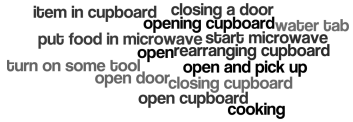


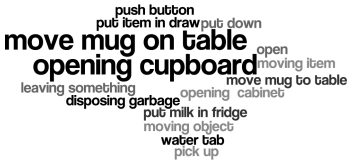
Figure 4: Segmentation test results.

(a) *Open door*



(b) *Close door*

(c) *Open Fridge*



(d) *Close Fridge*

(e) *Open Dishwasher*

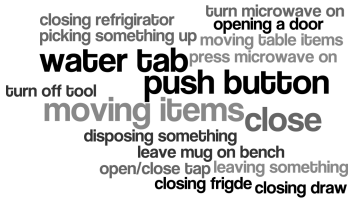
(f) *Close Dishwasher*

(g) *Open Drawer*

(h) *Close Drawer*

(i) *Clean Table*

(j) *Drink from Cup*



(k) *Toggle Switch*

Figure 5: Tag cloud of words used by testers in the open-ended experiment for each activity.

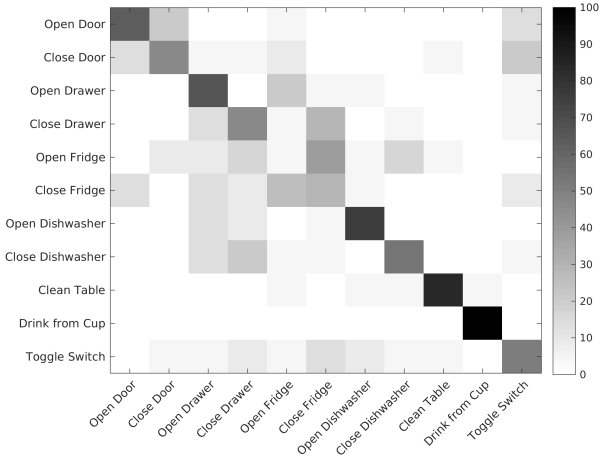


Figure 6: Confusion matrix of annotation choosen by users and actual labels.

est accuracy (4.2%). The main cause is likely the absence of any point of reference for the environment and the fact that the fridge was a small model. For this reason, the action of opening and closing can be easily confused with other actions applied to the same height, such as "Open a drawer". Moreover, the confusion between "Open/Close the door" is due to the lack of information about the direction of opening and closing of the door.

Finally, the partecipants reached an average accuracy of 56% in the controlled labelling experiment using our system.

Discussion

Our experiment revealed that a 3D human model can be used for activity annotation preserving the privacy of the user, but it would require some improvements.

About the segmentation, our tests showed that further analyses are required such as the capability of point out the duration of the action. Allow to identify the begin and the end of an activity can improve the accuracy in this step. It can be also important to specify to the user the granularity of the action. As we noticed during the experiment, the main trouble for the testers was: "What should I consider as an action to point out?". Answering this question can depend on the specific application scenario of each dataset: in some cases an action can be a simple gesture as "move the right arm up", "move the left hand down", etc. In others scenarios instead, it can be important to identify more complex actions as "make a sandwich", "prepare a coffee",etc.

Moreover, during our experiments, we pointed out a main issues in the lack of the environment in the scene. It can be difficult for the users to recognize the wide set of possible activities without knowing the position of the objects and of the furniture in the environment. A typical example of this is the confusion between the opening of the fridge and the opening of the drawer: these two movements appear similar when reproduced with a simple model such as our. This confusion between movements that appear similar is more observable in the "open-ended" annotation: in fact in this scenario it occurs that users identify correctly the movements (opening and closing), but it annotates the task with a different object whom those movements are applied to (the dishwasher mistaken with the oven).

Some improvements should be also applied to the model itself. In this first implementation, we used only basic solids to create the human figure: this brought some difficulties for the users to recognize actions made by short and limited movements. An example can be the toggling of the switch: in this case, the absence of the hands made it tricky to identify it. For this reason, we should explore whether a

more realistic human model could improve the accuracy of the annotation.

In order to improve the accuracy of the movements played by the model, a larger number of sensors can be a solution. In our tests the data animate only two parts of each upper limb and the torso, but the model has been developed to be animated with at maximum 12 sensors. The data from all these sensors can also be applied on the hands, on the legs and on the head. Furthermore, the software can be used with many different datasets passing specific parameters at start-up. The only requirement is that the dataset should contain IMU data for each body part the users want to animate. This can be a limitation: in fact, it can be difficult to use this system with those datasets already recorded and where the IMUs are placed only on few body parts, not allowing to the model to reproduce all the movements correctly. Instead, for those researchers that would use this system in the future, recording new datasets, it can be really a choice. Using our system, they can replace altogether the need for cameras. It means a saving both in costs and in time because researchers will not need neither equipments to record the videos nor additional time to preprocessing them.

It is the first time that annoation using a 3D model has been proposed. Even though an average accuracy of 56% can not be enough for ground truth, this system can be joined with algorithm of decision fusion (e.g.majority voting) and filters, to improve accuracy as already done in [8] for video annotation.

Conclusion

In this work we raise the need to create a privacy preserving annotation system. This in order to speed up the process of labelling dataset using cheap labour and crowdsourcing, where a video can not be used due to lack of anonymity of the user recorded in the video itself.

We study to which extent a 3D human model animated in a virtual environment using the data from inertial sensors can be used to annotate the dataset. To tests this we created a model and using a prior labelled dataset we animated it. We want to compare the annotation collected with our model and the actual labels of the dataset.

We developed three tests: a first one to study the capability of the user to recognize the activities done by the model and correctly segment the dataset. This is effectively the first step during the annotation process. The second experiment is studied to analyze to which extent the application of our annotation system in a "open-ended" scenario where the user can choose freely the label for each action. In the last test we investigate the accuracy of the annotation when a set of possible choices are given to the users, reaching an high level of truthfulness for some specific actions and manifold results for others.

From the tests, it appears a threefold result. The segmentation step requires furter analysis in order to better evaluate to which extent a 3D human model can be actually used for this task. The obtained results are not enough to give a strong positive judgment. The "open-ended" annotation can be used but only when dataset consist of actions that look very clear when reproduced by the model (e.g.drinking). For those actions where movements are limited and short, the accuracy with our system drops. Instead, using a "closed set" of annotations, our system allows users to reach an average accuracy of 56% also using very similar actions and with only 6 people. In this scenario, using more people and applying jointly decision fusion algorithms and filters for bad taggers, our system can be an actual

choice to annotate data preserving the privacy of the subject in the dataset.

REFERENCES

1. Amazon. 2005. Amazon Mechanical Turk. (2005). https://www.mturk.com/mturk/welcome, accessed 16/06/2016.
2. Akin Avci, Stephan Bosch, Mihai Marin-Perianu, Raluca Marin-Perianu, and Paul Havinga. 2010. Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey. In *Architecture of computing systems (ARCS), 2010 23rd international conference on*. VDE, 1–10.
3. Ling Bao and Stephen S Intille. 2004. Activity recognition from user-annotated acceleration data. In *Pervasive computing*. Springer, 1–17.
4. Andreas Bulling, Ulf Blanke, and Bernt Schiele. 2014. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys (CSUR)* 46, 3 (2014), 33.
5. J. Dai, J. Wu, B. Saghafi, J. Konrad, and P. Ishwar. 2015. Towards privacy-preserving activity recognition using extremely low temporal and spatial resolution cameras. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 68–76. DOI: http://dx.doi.org/10.1109/CVPRW.2015.7301356
6. Susumu Harada, Jonathan Lester, Kayur Patel, T Scott Saponas, James Fogarty, James A Landay, and Jacob O Wobbrock. 2008. VoiceLabel: using speech to label mobile sensor data. In *Proceedings of the 10th international conference on Multimodal interfaces*. ACM, 69–76.
7. The jME core team. 2016. jMonkeyEngine. (2016). http://jmonkeyengine.org/.
8. Long-Van Nguyen-Dinh, Cédric Waldburger, Daniel Roggen, and Gerhard Tröster. 2013. Tagging human activities in video by crowdsourcing. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*. ACM, 263–270.
9. Gabriel Parent and Maxine Eskenazi. 2011. Speaking to the Crowd: Looking at Past Achievements in Using Crowdsourcing for Speech and Predicting Future Challenges.. In *INTERSPEECH*. Citeseer, 3037–3040.
10. Daniel Roggen, Alberto Calatroni, Mirco Rossi, Thomas Holleczek, Kilian Förster, Gerhard Tröster, Paul Lukowicz, David Bannach, Gerald Pirkl, Alois Ferscha, and others. 2010. Collecting complex activity datasets in highly rich networked sensor environments. In *Networked Sensing Systems (INSS), 2010 Seventh International Conference on*. IEEE, 233–240.
11. Tim Van Kasteren, Athanasios Noulas, Gwenn Englebienne, and Ben Kröse. 2008. Accurate activity recognition in a home setting. In *Proceedings of the 10th international conference on Ubiquitous computing*. ACM, 1–9.
12. Carl Vondrick, Donald Patterson, and Deva Ramanan. 2013. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision* 101, 1 (2013), 184–204.
13. Aobo Wang, Cong Duy Vu Hoang, and Min-Yen Kan. 2013. Perspectives on crowdsourcing annotations for natural language processing. *Language resources and evaluation* 47, 1 (2013), 9–31.
14. XSens. 2000. MTx 3D Tracker. (2000). https://www.xsens.com/products/mtx/